Computer and Information Science; Vol. 14, No. 4; 2021
ISSN 1913-8989 E-ISSN 1913-8997
Published by Canadian Center of Science and Education

Automatic Identification and Filtration of COVID-19 Misinformation

Paras Gulati[1], Abiodun Adeyinka. O.[2], & Saritha Ramkumar[3]

[1] New York University, New York, United States

[2] Bowen University Iwo, Osun State, Nigeria

[3] University of North Carolina, Charlotte, United States

Correspondence: Abiodun Adeyinka, Osun State, Nigeria.

Received: August 24, 2021 Accepted: October 11, 2021 Online Published: October 26, 2021

doi:10.5539/cis.v14n4p57 URL: https://doi.org/10.5539/cis.v14n4p57

Abstract

The rapid spread of online fake news through some media platforms has increased over the last decade. Misinformation and disinformation of any kind is extensively propagated through social media platforms, some of the popular ones are Facebook and Twitter. With the present global pandemic ravaging the world and killing hundreds of thousands, getting fake news from these social media platforms can exacerbate the situation. Unfortunately, there has beena lot of misinformation and disinformation on COVID-19 virus implications of which has been disastrous for variouspeople, countries, and economies. The right information is crucial in the fight against this pandemic and, in this age of data explosion, where TBs of data is generated every minute, near real time identification and tagging of misinformation is quintessential to minimize its consequences. In this paper, the authors use Natural Language Processing (NLP) based two-step approach to classify a tweet to be a potentially misinforming one or not. Firstly, COVID -19 tagged tweets were filtered based on the presence of keywords formulated from the list of common misinformation spread around the virus. Secondly, a deep neural network (RNN) trained on openly available real and fake news dataset was used to predict if the keyword filtered tweets were factual or misinformed.

Keywords: COVID-19, Corona virus, Tweets, Infodemic, Fake news, social media, Pandemic, SARS-CoV-2

1. Introduction

Social media is a significant conduit for news and information in the modern media environment (P. Sharma and Kaur, 2017), with one in three people in the world engaging in social media, and two thirds of those on the internet usingit (Ortiz-Ospina, 2019). There were 255 million Monthly active Twitter users in February 2014 during the start ofthe Ebola outbreak (Twitter, 2014). But, in February 2020, during the start of the Covid-19 outbreak, twitter reports 166million daily active users (Twitter, 2020). This shows the magnitude of reach any information or misinformation can have in this one social media platform. As much as it provesto be the novel means of connecting people and disseminating information, it has also proved to provide too much information and misinformation, causing hysteria, mentaldistress, self-harm and in few cases, suicides (Rosenberg, Syed, and Rezaie, 2020).

Misinformation can be defined as a claim of fact that is currently false due to lack of scientific evidence (Chou, Oh, and Klein, 2018). Also, in the study by Kouzy et. al. (2020), authors states that tweet quality (misinformation vs correct information) did not differ based on the number of likesor retweets, indicating that misinformation is as likely to spread and engage users as the truth (Kouzy et al., 2020). Also, more than 40% of the total tweets includes an URL, which presumably indicate authenticity, indicate that only 0.4% of those are from very credible health sources likethe CDC and WHO (Singh et al., 2020). As more and more people live in isolation, fearing the risk of outbreak, they are more prone to probe more information about the disease, which demands authenticity. This need has been acknowledged by the World Health Organization, which haspartnered with several social media platforms and seven major tech companies—namely, Facebook, Google, LinkedIn, Microsoft, Reddit, Twitter, and YouTube—that agreed to stamp out fraud and misinformation, and to promote critical updates from healthcare agencies (Statt, 2020). Identifying hoaxes and rumors from online platforms and segregating them from scientific information's can be achieved through NLP and advanced text analytics. Applying principles and learning from the algorithms and methodologies used by online companies like Amazon, reddit etc. to detect fake accounts and fake reviews could serve as a substantial meansto identify misinformation

(Shu et al., 2017).

In this paper, we discuss our approach to classify Covid-

19 related misinformation in Twitter using a Recurrent Neural Network model that classify Real and fake news. Wecollected tweets related to Covid 19 and do an advanced keyword-based filtration and categorization based on thepresence of keywords related to the common misinforma- tion. The newly filtered tweets are now classified as factual or misinformed using the text classification model based on real and fake news datasets.

2. Related Works

The research by Singh et al. (2020), summarises all Covid-19 related tweets into eight predefined themes using a word to theme matching approach- Economy (example words:Market, stocks, futures), Emotion (example words: Fear, joke, hope), Healthcare/Illness/Virus (example words: Pa- tients, coronavirus, infected, vaccine, tested, SARS), Global Nature (example words :Pandemic, international, China, Italy, travel), Information Providers (example words: Me- dia, CDC, WHO, experts), Social (example words: Family, friends, community), Government/ Government Response (example words: Trump, senator, lockdown), Individual Concerns/Strategies(example words: Disinfect, wash, facemasks)(Singh et al., 2020).

The research work by Sharma (2020) approaches the problem in multiple angles. They identify misinformation based on information cascades (the source tweets and the propagation information from the cluster of the retweet graph) and analyses the degree of falsehoods and varyingor deliberate intents, thereby classifying them into four categories - Unreliable (false, questionable, rumors and mis-leading news), Conspiracy (conspiracy theories and scientif- ically dubious news), Click-bait (exaggerated or misleading headlines to attract attention) and Political/Biased (written insupport of a particular political orientation). They analyzed the geographical spread for each of these categories and theirsentiment (K. Sharma et al., 2020).

Whereas the research by Shahi (2020) follows a two-step approach of first identifying the accounts involved in the spread of misinformation followed by analysis (Shahi, Dirkson, and Majchrzak, 2020). Accounts are categorizedas: 1) ‘bot’ accounts, 2) accounts associated with brandsand lastly, 3) their popularity or follower count. Informationdiffusion is analyzed by using the speed of retweets as a proxy for the speed of propagation, which was followed by content analysis using hashtags, emojis and distinctive terms.

3. Methodology

A. Data Collection

COVID-19 related English tweets from the United States of America were obtained within the period of March 20, 2020, to April 30, 2020. There were 4 million tweets obtained; out of which more than 77% were geotagged. Table I summarizes the stats and Table 2 lists the hashtags that were used.

Table 1. Tweets Statistics

March 20 - April 30	**Count**
Total Tweets	4199878
Total Tweets with Geo Information	3086549(77.77%)
Total Number of Accounts	3372189

B. Formulating a keyword corpus of popular misinformation spread around COVID -19

IDeaS Center and CASOS Center released a curated list ofstories containing inaccurate information regarding COVID-19, categorized into six categories- cure, nature of virus, con-spiracy theories, emergencies, misbehavior, and good news (Carley, 2020). We extracted keywords from each of these categories using RAKE library (Python implementation of the Rapid Automatic Keyword Extraction algorithm using NLTK) and stored them in separate files for next steps.

Table 2. Keywords (Hashtags) Used For Tweets Extraction

S. No.	Keywords
1	COVID19
2	CoronavirusPandemic
3	COVID-19
4	2019nCoV
5	CoronaOutbreak
6	coronavirus
7	WuhanVirus
8	wuhan
9	pneumonia
10	pneumonie
11	neumonia
12	lungenentzündung
13	covid19

3. Summarizes the List of Keywords or Phrases Extracted in Each Category.

C. Tweets Pre-processing and Categorization

The scraped data from twitter was stored in a file and cleaned by removing html tags using Beautiful Soup library.PhraseMatcher API from Spacy was used to match the previously extracted keywords with the tweets, indicating a tweet that may contain misinformation related to one ofthe categories from table 3.

D. Classification Model

We used about 35,918 news articles to train on the model and 8,980 to test the model. The minimum length of the article in the training dataset is 32 words and maximum is 51,894. The median length of articles is 2,269 words. We truncated the articles to a maximum length of 300 words.We used top 10,000 words and tokenized them. The smaller articles were padded with zeros at the end.

All the words in the text sentences are converted to low dimensional vectors. These word vectors are then stacked to create an embedding matrix: $E_w \epsilon R^{d\times|v|}$. Here d signifies the dimension of the vector and v signifies the vocabulary size. Each word is mapped into 100 dimensional vector using pre-trained GloVe embeddings (Pennington, Socher, and Manning, 2014) to represent the words into higher dimensional vector space.

Next, these word embedding are sent into Long Short Term Memory (LSTM) cells (Graves, 2012). Each LSTM cells is fed with word vector at the current time step aswell as the output from the previous LSTM cells. In this fashion, the LSTM layer learns the patterns that representa fake news or a fact. LSTM has an advantage over vanilla RNN that it tackles the vanishing gradient problem, making it widely popular choice for NLP applications that requires long range dependencies (Graves, Mohamed, and Hinton, 2013), (Rao et al., 2018). Since, it is a binary classification we used binary crossentropy loss function:

Table 3. Sample Keywords From Each Category

Category	Keywords/Phrases
Cure	hot bath prevents, sesame oil prevents, drinking hot water mustard oil kills, sheep head',drinking cow urine, drinkingcorona beer, cocaine kills, propolis cures, neem tree leaves, miracle mineral supplement
Nature of Virus	children cannot catch corona virus, dead birds due, turks area immune, 19 breeds rapidly, poisoning, unborn child,toilet paper, pregnant women, normal flu, mutating faster, mosquito bite
Conspiracy Theory	thailand carried, doha paid billions, vaccine called interferon, military world games, us military brought, franceinvented ,19 since 2015, egypt gave china, soap companies, selling sars
Emergency	invoke martial law, infected toilet paper, india deploys army, hungary forcing men, human interaction needs, caughtwandering outside, nationwide mandatory quarantine, committing suicide due
Misbehavior	spread corona virus, muslim owned restaurants, italy commits suicide, finance cultural projects, muslims amidst,loses family, huge stashes, criminals handing, commit robbery
Good News	sri lankan sambar deer graze, pandemic caused venice, yala park beach, ciudad real spain, deers walking, villagedue

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}(y_i log(p_i)) + (1-y_i)log(1-p_i)$$

Here y_i is the target value and p_i is the predicted value of the model. N denotes the number of scalar values in the output of the model.

In order to predict the class, we used a sigmoid activation function. Sigmoid is a non-linear function that squeezes the input values between 0 and 1.

$$f(x) = \frac{1}{1 + e^{-x}}$$

The model architecture had the following layers in sequential order: input layer, embedding layer, LSTM layer with 128 hidden parameters, dropout layer with dropout value 0.2, LSTM layer with 64 hidden parameters, dropout layer with dropout value 0.2, fully connected layer with 32 units and ReLU activation function, and lastly fully connected layer with 1 unit and sigmoid activation function. We used adam optimizer to compile the model and the model architectureis presented in figure 1.

The model, when trained on label data of true and fake news articles gives and accuracy of 99% on test data. The Precision, recall and F1 score are mentioned in table V.

E. Making Predictions on Tweets

Once the tweets were filtered as mentioned in III-C, these tweets were then tokenized, padded and passed through the fake news classification model, described in III-D, toget a score ranging between 0 and 1. We can then usean appropriate threshold value to tag the misinformationcarrying tweets and thus counter the spread of covid-19related misinformation.

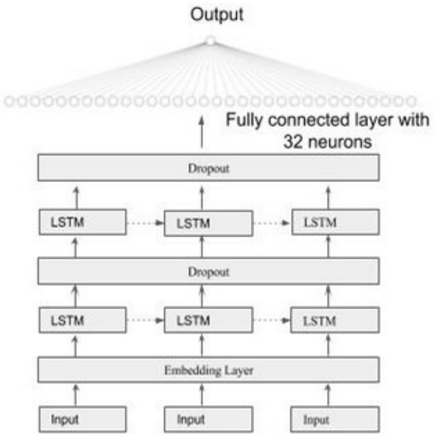


Figure 1. Model Architecture

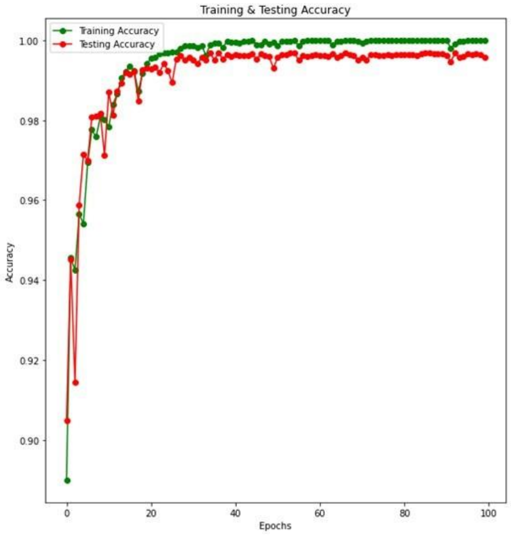


Figure 2. Training and Test Accuracies

Table 4. Few Examples of Tweets after Keyword Matching

Tweet	**Location**	**Cure**	**Nature of Virus**	**Conspiracy Theory**	**Emergency Measure**	**Good News**	**Misbehavior**
@thekiranbedi Good morning. Stay home, keep social distance, no handshakes, no sneezing, wash hands, use sanitizer, stop the spread of corona virus, stay safe.https://t.co/BLStVenT5T	Kentucky, USA	[’wash’]		[’stay’, ’home’, ’spread’, ’stay’]	[’stay home’, ’sneezing’,’stop’, ’spread’]		
4 myths about COVID-19 debunked. Cold weather or snow won’t kill the virus. Taking a hot bath won’t prevent corona virus. The virus cannot be transmitted through mosquitos bites. Antibiotics aren’t effective in treating/preventing COVID-19. Self isolate and wash your hands! https://t.co/r9joXcJ9m3	Washington, DC	[’cold weather’, ’snow’, ’kill’,’taking’, ’prevent’, ’antibiotics’ , ’effective’, ’treating’, ’preventing’, ’wash’]	[’cold’, ’transmit-ted’]		[’kill’, ’iso-late’]		[’prevent’]
Please wash your hands with soap and water regularly or use an alcohol based hand san-itizer where soap and water isn’t available. Also practice social distancing so we can put an end to Corona virus ASAP. #FightCovid19	Lagos, Nige-ria.Texas, USA	[’wash’, ’hand sanitizer’]			[’social dis-tancing’]	[’water’, ’water’, ’social distancing’]	
@JayCridlin My favorite drink, a Coronarita. Its not intended to make light of what we are all going through right now ... but I just cant believe people actually stopped drinking Corona beer! We are VERY lucky to have a pool, and things could be mad	Tampa Bay	[’make’, ’drinking’, ’drinking corona beer’]	[’people’]	[’worse’, ’us’]	[’going’, ’people’, ’stopped’,’us’]	[’drink’, ’people’]	
@CNN @lisaling @AndrewYang Why is assimilation so difficult for Chinese people and when do they support non-Chinese businesses? In my community, Chinese predominantly if not exclusively support their own yet at the start of corona people were racist	CambidgeBoston		[’people’, ’people’]	[’chinese’,’chinese’, ’chinese’]	[’people’, ’people’, ’stopped’,’going’]	[’people’, ’people’]	
woww I’ve been drinking hot water +lime and i just found out it does not help against corona nice.	New Jersey	[’drinking’, ’drinking hot water’]			[’water’]		
I don’t masterbate anymore.I wash my hands and touch my face for 10 minutes. . . . #Corona #coronavirus #Coron-aAlert	New York City	[’wash’, ’10 minutes’]					
Gays get stds and still have no problem inviting people overfor fun...if you think they’ll stop with covid you are sadly mistaken just wash your hands wtf lmao	Scottsdale,AZ	[’get’, ’wash’]	[’people’]		[’people’, ’stop’]		[’people’, ’covid’]

Table 5. Results For Fake News Detection Model

Class	Precision	Recall	F1 Score
Fake News	0.9936	0.9981	0.9959
Factual News	0.9979	0.9930	0.9954
Macro Avg	0.9958	0.9955	0.9956
Weighted Avg	0.9957	0.9957	0.9957

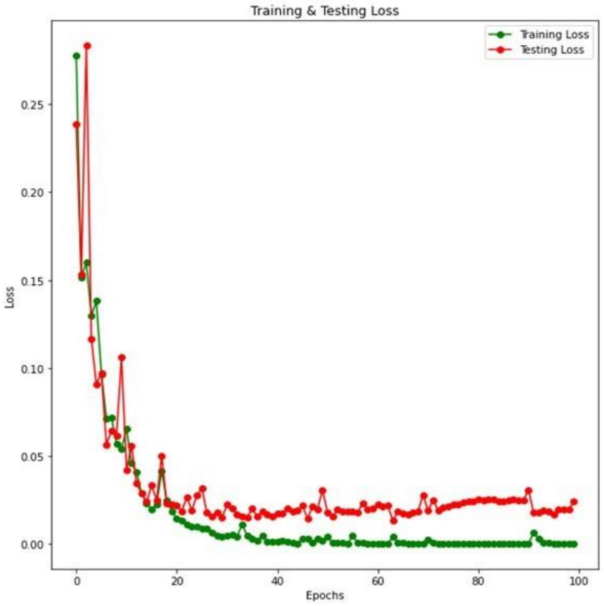
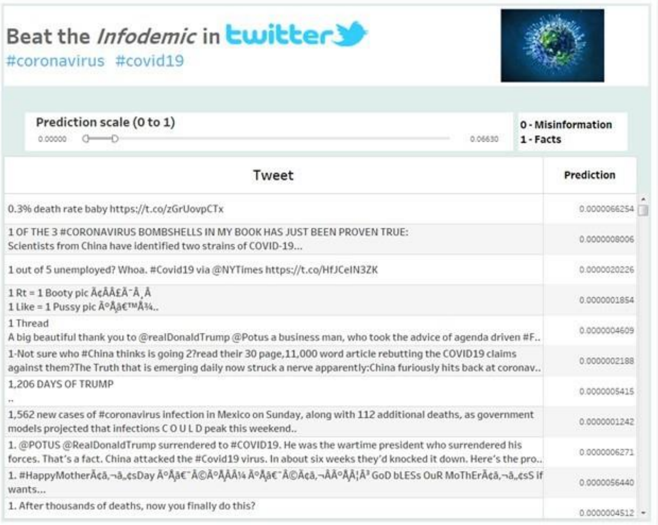


Figure 3. Training and Test Loss

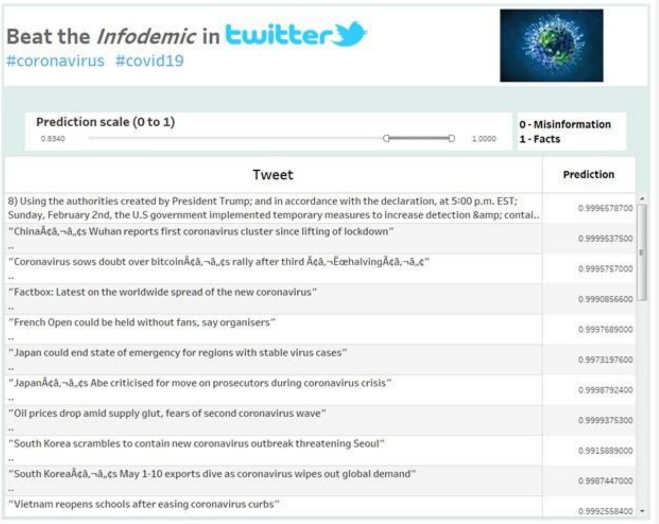
4. Results

When we set the threshold of the prediction to a higher number, 0.8, we get genuine tweets and when we set the threshold to a lower value, 0.2, then we get the misinforming tweets related to COVID-19; examples can be viewed in the tableau dashboard: Tableau Dashboard.



Tweet	Prediction
0.3% death rate baby https://t.co/zGrUovpCTx	0.0000066254
1 OF THE 3 #CORONAVIRUS BOMBSHELLS IN MY BOOK HAS JUST BEEN PROVEN TRUE: Scientists from China have identified two strains of COVID-19...	0.0000008006
1 out of 5 unemployed? Whoa. #Covid19 via @NYTimes https://t.co/HfJCeIN3ZK	0.0000020226
1 Rt = 1 Booty pic Ã¢ÂÂ£Ã¯Â¸Â 1 Like = 1 Pussy pic Ã°Å¸â€™Â¾..	0.0000001854
1 Thread A big beautiful thank you to @realDonaldTrump @Potus a business man, who took the advice of agenda driven #F..	0.0000004609
1-Not sure who #China thinks is going 2?read their 30 page,11,000 word article rebutting the COVID19 claims against them?The Truth that is emerging daily now struck a nerve apparently:China furiously hits back at coronav..	0.0000002188
1,206 DAYS OF TRUMP ..	0.0000005415
1,562 new cases of #coronavirus infection in Mexico on Sunday, along with 112 additional deaths, as government models projected that infections C O U L D peak this weekend..	0.0000001242
1. @POTUS @RealDonaldTrump surrendered to #COVID19. He was the wartime president who surrendered his forces. That's a fact. China attacked the #Covid19 virus. In about six weeks they'd knocked it down. Here's the pro..	0.0000006271
1. #HappyMotherÃ¢â‚¬â„¢sDay Ã°Å¸â€˜Â©Ã°Å¸Â¼ Ã°Å¸â€˜Â©Ã¢â‚¬ÂÃ°Å¸Â¦Â³ GoD bLESs OuR MoThErÃ¢â‚¬â„¢sS if wants...	0.0000056440
1. After thousands of deaths, now you finally do this?	0.0000004512

Figure 4. Misinformation/Non-Facts Tweets



Beat the *Infodemic* in twitter

#coronavirus #covid19

Prediction scale (0 to 1)

0.8340 1.0000

0 - Misinformation

1 - Facts

Tweet	Prediction
8) Using the authorities created by President Trump; and in accordance with the declaration, at 5:00 p.m. EST; Sunday, February 2nd, the U.S government implemented temporary measures to increase detection &amp; contai..	0.9996578700
"ChinaÃ¢â‚¬â„¢s Wuhan reports first coronavirus cluster since lifting of lockdown" ..	0.9999537500
"Coronavirus sows doubt over bitcoinÃ¢â‚¬â„¢s rally after third Ã¢â‚¬ËœhalvingÃ¢â‚¬â„¢" ..	0.9995757000
"Factbox: Latest on the worldwide spread of the new coronavirus" ..	0.9990856600
"French Open could be held without fans, say organisers" ..	0.9997689000
"Japan could end state of emergency for regions with stable virus cases" ..	0.9973197600
"JapanÃ¢â‚¬â„¢s Abe criticised for move on prosecutors during coronavirus crisis" ..	0.9998792400
"Oil prices drop amid supply glut, fears of second coronavirus wave" ..	0.9999375300
"South Korea scrambles to contain new coronavirus outbreak threatening Seoul" ..	0.9915889000
"South KoreaÃ¢â‚¬â„¢s May 1-10 exports dive as coronavirus wipes out global demand" ..	0.9987447000
"Vietnam reopens schools after easing coronavirus curbs"	0.9992558400

Figure 5. Legitimate Information/Facts Tweets

5. Conclusion

We observed that the LSTM based sequential model can learn the tone, sentiment, and genuineness of a text when trained on a labelled data of fake news and can classifythe misinformation from the real information with high accuracy. In the current study, we attempted to segregate thecovid-19 related misinformation from 4M tweets collected over 40 days period using a model which is pre-trained on fake news dataset. We can see from the tableau dashboard that on a higher threshold (0.8 in this case), we get genuine tweets that do not carry any misinformation. On the other side, when we use a lower threshold (0.2), it indicated misinformation containing tweets.

6. Discussions and Limitations

This was an attempt to classify misinformation relatedto COVID-19 and counter the spread of it. We extracted tweets that indicate a possible misinformation theory based on keywords and used a model that was trained on a labeled dataset of true and fake news articles to classify the tweets. The misinformation tweets identified by the model can be used to further train the model to identify the patterns in misinforming tweets related to COVID-19. In this waythe model can learn new patterns and language related to COVID-19. Another better approach can be to check the misinforming tweets predicted by the model and after fact- checking, we feed the cleaner data to the model for training. The current model is based on LSTM approach and uses twitter data of about 4M tweets collected over a period of about a month. In the future work of this paper, we will collect data over wider span of time for our analysis and explore more advanced NLP models including transformer- based models for comparisons with the current models.

References

Carley, K. M. (2020). *List of Known Misinforma-tion and Disinformation Regarding Corona Virus in So- cial Media*. Retrieved from https://www.cmu.edu/ideas-social-cybersecurity/research/misinformation-and-disinformation-4-16-2020.pdf

Graves, A. (2012). Long short-term memory. In *Super-vised sequence labelling with recurrent neural networks*. Springer, pp. 37-45. https://doi.org/10.1007/978-3-642-24797-2_4

Graves, A., Abdel-rahman, M., & Geoffrey, H. T. (2013). Speech recognition with deep recurrent neu- ral networks. In *2013 IEEE international conference on acoustics, speech and signal processing*. Ieee, pp. 6645-6649. https://doi.org/10.1109/ICASSP.2013.6638947

Kouzy, R. et al. (2020). Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter. *Cureus, 12*. https://doi.org/10.7759/cureus.7255

Pennington, J., Richard, S., & Christopher, D. M. (2014). Glove: Global vectors for word repre- sentation. In *Proceedings of the 2014 conference on em- pirical methods in natural language processing (EMNLP),* pp. 1532-1543. https://doi.org/10.3115/v1/D14-1162

Rao, G. Z. et al. (2018). LSTM with sentence rep- resentations for document-level sentiment classification. *Neurocomputing, 308*, 49-57. https://doi.org/10.1016/j.neucom.2018.04.045

Rosenberg, H., Shahbaz, S., & Salim, R. (2020). The Twitter pandemic: The critical role of Twitter in the dissemination of medical information and misinfor- mation during the COVID-19 pandemic. *CJEM, 22*(4), 418-421. https://doi.org/10.1017/cem.2020.361

Shahi, G. K., Anne, D., & Tim, A. M. J. (2020). *An Exploratory Study of COVID-19 Mis- information on Twitter*. arXiv: 2005.05710 [cs.SI].

Sharma, K. et al. (2020). *COVID-19 on Social Me-dia: Analyzing Misinformation in Twitter Conversations*. arXiv: 2003.12309 [cs.SI]. https://doi.org/10.1016/j.osnem.2020.100104

Sharma, P., & Pankaj, D. K. (2017). Effective-ness of web-based social sensing in health information dissemination—A review. *Telematics and Informat- ics, 34*(1), 194-219. https://doi.org/10.1016/j.tele.2016.04.012

Shu, K. et al. (2017). *Fake News Detection on Social Media: A Data Mining Perspective*. arXiv: 1708.01967 [cs.SI].

Statt, N. (2020). Major tech platforms say they're 'jointlycombating fraud and misinformation' about COVID-19. *The Verge, 3*. Retrieved from https://www.theverge.com/2020/3/16/21182726/coronavirus-covid-19-facebook-google-twitter-youtube-joint-effort-misinformation-fraud

Twitter. (Apr. 2014). *Twitter Reports First Quarter 2014 Results.* Retrieved from https://s22.q4cdn.com/826641620/files/docfinancials/2014/q1/2014Q1Earnings Release.pdf

Chou, W. Y. S., April, O., & William, M. P. K. (Dec. 2018). Addressing Health-Related Misinformationon Social Media. *JAMA, 320*(23), 2417-2418. https://doi.org/10.1001/jama.2018.16865

Ortiz-Ospina, E. (Sept. 2019). *The rise of social media. Our World in Data*. Retrieved from https://ourworldindata. org/rise-of-social-media

Singh, L. et al. (Mar. 2020). *A first look at COVID-19 information and misinformation sharing on Twitter. eng. ArXiv,* arXiv:2003.13907v1. Retrieved from https://pubmed.ncbi.nlm.nih.gov/32550244

– (Apr. 2020). *Twitter Announces First Quarter 2020 Results.* Retrieved from https://s22.q4cdn.com/826641620/files/doc financials/2020/q1/Q1-2020-Earnings-Press-Release.pdf

–

Copyrights